Table of Contents

Section: Ordination analysis

Supplementary variables (unconstrained ordination)

Theory R functions Examples Exercise

The ecological meaning of the axes in unconstrained ordination can be interpreted by relating the sample scores on these axes to external supplementary variables (usually measured or estimated environmental variables). This relationship can be done either by correlating the supplementary variable to the first two or few main axes using Pearson's correlation coefficient or by regressing the supplementary on the sample scores of selected ordination axes using (weighted) multiple regression. The **correlation** method is more intuitive to understand, but the application is limited only to linear ordination methods, while the **(weighted) multiple regression** is less intuitive, but more general, applicable to both linear and unimodal ordination methods. The results are often used to project supplementary variables passively onto the ordination diagram while reporting the strength of the relationship with ordination axes (correlation coefficient in the case of correlation, r^2 in the case of multiple regression) and possibly also the test of significance. There is a difference between linear and unimodal ordination method; while in the linear method all samples have the same weight, in the unimodal method the sample weight (its importance in the analysis) is proportional to the sum of species abundances in this sample. This has to be reflected when the supplementary variables are related to ordination axes, and the weights need to be included in the calculation (that's why weighted multiple regression is used).

Correlation of supplementary variable with selected ordination axes

The method is illustrated on a simple example of tb-PCA, unconstrained linear ordination method (principal component analysis applied on Hellinger-transformed species composition data). It consists of the following steps:

- species composition data (Fig. 1a) are used to calculate sample scores on ordination axes (Fig. 1b);
- select axes you want to interpret (often first two axes or few more, rarely more than three), and calculate Pearson's correlation coefficient between samples scores on each selected PCA axis (Fig. 1b) and each of the environmental variable (Fig. 1c);
- In the case of quantitative supplementary variables, the correlation coefficient (Fig. 1d) can be projected into the space of ordination diagram defined by selected ordination axes using the vectors (Fig. 1e), starting at coordinates [0,0] (the centre of the diagram) and pointing toward [r_1, r_2]. The angle and the direction of the vector is dependent on the value and sign of the respective correlation coefficient (in our example, the pH is relatively weakly positively correlated with PCA1 (correlation coefficient r_1) and rather strongly and negatively with PCA2 (r_2); as a result, the vector arrow is pointing toward the right bottom corner of the diagram, and is more tightly related to the second than to the first PCA axis.
- Resulting ordination diagram (Fig. 1f) usually apart to passively projected supplementary variables includes also sample or species scores (biplot) or both (triplot).

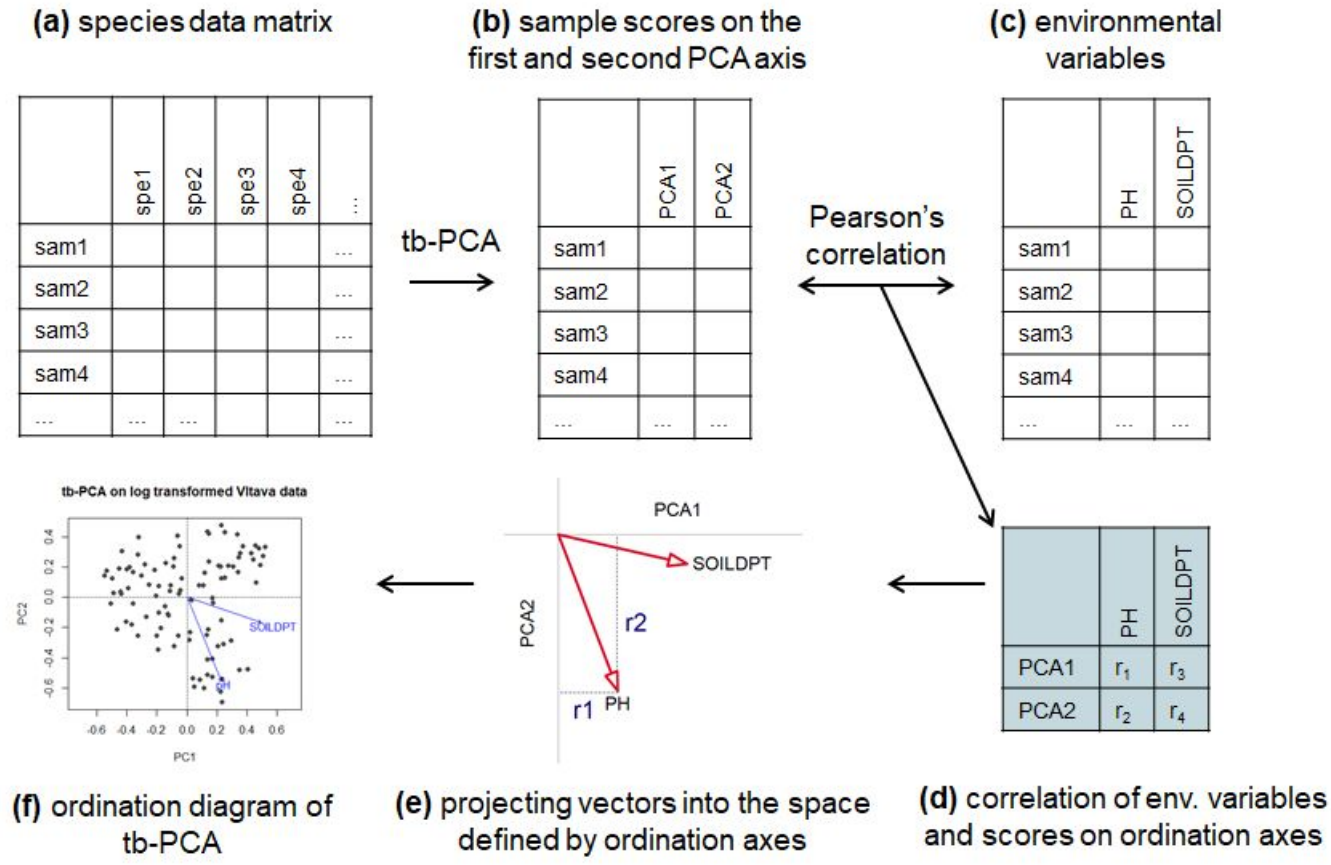


Figure 1

(Weighted) multiple regression of supplementary variables on selected ordination axes

In this analysis, each supplementary variable is used as a dependent variable in the multiple regression with selected ordination axes as explanatory variables, following this equation (in the case of two ordination axes):

$$env = b_0 + b_1 score_1 + b_2 score_2 + \epsilon$$

where b_1 and b_2 are regression coefficients, and $score_1$, $score_2$ are sample scores on the ordination axes. Resulting regression coefficients, after being normalized and multiplied by the regression's coefficient of determination (r^2) are equal to correlation coefficients described in the section above and can be used to plot the vectors onto ordination diagram. This equivalence, however, works only for linear regression (PCA, tb-PCA); in the case of unimodal ordination (CA, DCA), additional weights need to be applied. Each regression can be also tested, using the Monte Carlo permutation test, to ensure whether the fit of the supplementary variable on ordination axes is significant and hence worth to interpret and project onto the diagram.

Some more technical details. Before the analysis, both dependent and explanatory variables are each centered (to have zero mean), but not standardized; in case of unimodal ordination, the weights are calculated as row sums of species abundances in each sample, and these weights are used for the centering of all variables included in the multiple regression. Calculated b_0 (intercept) is zero, because all variables are centred. The regression coefficients b_1 and b_2 need to be normalized, so as their sum of squares equals to one. The normalized values c_2 and c_2, respectively, can be calculated as

$c_1 = b_1 \times k$, and $c_2 = b_2 \times k$, where *k* is the scaling factor, k = 1/sqrt (b_1^2 + b_1^2)). The c_1 and c_2 are called cosines, and if multiplied by r^2 from the above multiple regression, they are equal to r_1 and r_2 calculated using Pearson's correlation above (but only in case of linear regression without using weights, as mentioned above).

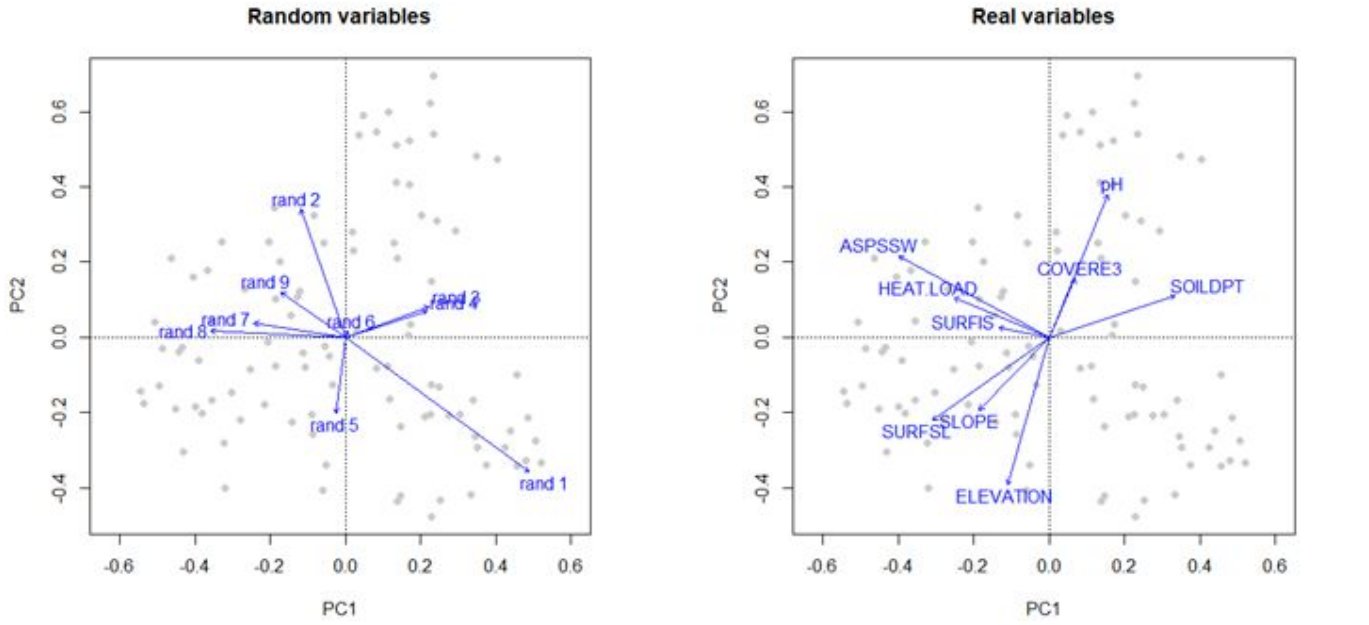
The results in the case of the two environmental variables from the example above are summarised in Table 1 (output of `vegan::envfit` function). The values in the columns PC1 and PC2 are not correlation coefficients, but cosines (see above; the sum of their squares equals to one). The column r^2 contains coefficient of determination for (weighted) multiple regression for each environmental variable; the higher is this values, the more important is the environmental variable to explain variation in the given number of ordination axes. The Pr column is the result of the Monte Carlo permutation test; it is perhaps better to use only significant variables for interpretation and plotting of the ordination diagram. Note that since the test is permutational, the minimum P-value depends on the number of permutations used, following the relationship P_{min}=1/(number_of_permutations + 1).

	PC1	**PC2**	**r^2**	**Pr(>r)**
pH	0.38093	-0.9246	0.3885	0.001***
SOILDPT	0.94892	-0.3155	0.289	0.001***

Table 1: Output of mutliple regression of two supplementary variables (pH and SOILDP) on first two ordination axes (tb-PCA in this case). See text for details.

Why to test the relationship of supplementary variables to ordination axes

Even if the supplementary variables are just a set of random variables, if projected onto the ordination diagram, they may appear as good to interpret (Fig. 2). The absolute length of the variable vector is dependent on the r^2 (in case of using multiple regression), and variables with higher r^2 have a longer vector. However, when projected onto the ordination diagram, the lengths of the vectors are rescaled relatively to the length of the variable with the highest r^2. If all variables are random, they still have non-zero r^2 (although quite low), and some will have higher r^2 than the others - this variable will appear with the longest vector, and all others will be scaled relative to it. This is why it is important to check the table of regression results. The absolute values of r^2 are difficult to interpret since they are influenced by the number of samples in the analysis (r^2 decreases with the increasing number of samples). This is why the statistical test can help, helping to figure out which variables have r^2 high enough to use for interpretation.



	PC1	PC2	r2	Pr(>r)	
rand 1	0.80832	-0.58874	0.0652	0.038	*
rand 2	-0.33099	0.94364	0.0232	0.319	
rand 3	0.93738	0.34831	0.0099	0.633	
rand 4	0.95066	0.31023	0.0093	0.645	
rand 5	-0.11882	-0.99292	0.0074	0.715	
rand 6	0.31065	0.95052	0.0001	0.999	
rand 7	-0.98738	0.15839	0.0109	0.602	
rand 8	-0.99894	0.04607	0.0228	0.330	
rand 9	-0.81692	0.57674	0.0078	0.705	

	PC1	PC2	r2	Pr(>r)	
ELEVATION	-0.27253	-0.96215	0.3836	0.001	***
SLOPE	-0.69341	-0.72055	0.1654	0.001	***
ASPSSW	-0.87822	0.47826	0.4769	0.001	***
HEAT.LOAD	-0.92198	0.38724	0.1728	0.001	***
SURFSL	-0.81522	-0.57915	0.3322	0.001	***
SURFIS	-0.98261	0.18569	0.0435	0.113	
SOILDPT	0.94893	0.31549	0.2891	0.001	***
pH	0.38093	0.92460	0.3885	0.001	***
COVERE3	0.40200	0.91564	0.0690	0.027	*

Figure 2: Nine randomly generated (left panel) and real (right panel) supplementary variables projected onto the ordination diagram. The results of multiple regression with the Monte Carlo permutation test are provided below. If we do not inspect the numeric results, we may be tempted to interpret the variables in the ordination diagram even if they are in fact random numbers. The tables show that, indeed, in case of the random variables all (but one) variables are insignificant, and r<sup

To test the significance of the relationship between a supplementary variable and ordination axes does not make sense in case that supplementary variables are in some way derived from the same species composition matrix as used in the ordination itself. This is e.g. the case when supplementary variables are species richness/diversity, or community weighted means (CWM) of species traits, Ellenberg-type indicator values or other species attributes. Since the null hypothesis tested is “there is no relationship between supplementary variable and samples scores on ordination axes”, it is easy to reject at the moment when these are in fact derived from the same information (species composition matrix). The solution is not to test the significance, or, in the case of CWM, use the modified permutation test based on permuting species attributes (see more in Zelený & Schaffers 2012 and Zelený 2018, and use the function `envfit_cwm` available in weimea package if you wish to apply the modified permutation test).

Multiple testing issue and available corrections

The more tests of significance we are doing, the higher is the chance to observe the significant result, even if the null hypothesis is true (no relationship). This rule is called *multiple testing issue* and can be illustrated in a simple example. I generated two random variables with normal distribution, calculated their regression, and tested it (using parametric F-test). One would expect that the test will not return a significant result since the variables are generated randomly. But if I repeat this 100

times (Fig. 3), you can see that some of the results turn to be significant. The proportion of significant results depends on the threshold value you use to deem result significant; e.g., if you consider as significant results with P-value lower than 5% (alpha = 0.05), then about 5% of the tests may appear as significant even though the variables are random (Type I error). Or, put in another way, the probability that at least one of the tests will be significant at P < alpha can be calculated $1 - (1 - m)^{alpha}$, which is called *family-wise Type I error rate* - the probability we are conducting Type I error rate if we interpret the results of multiple tests without any correction.

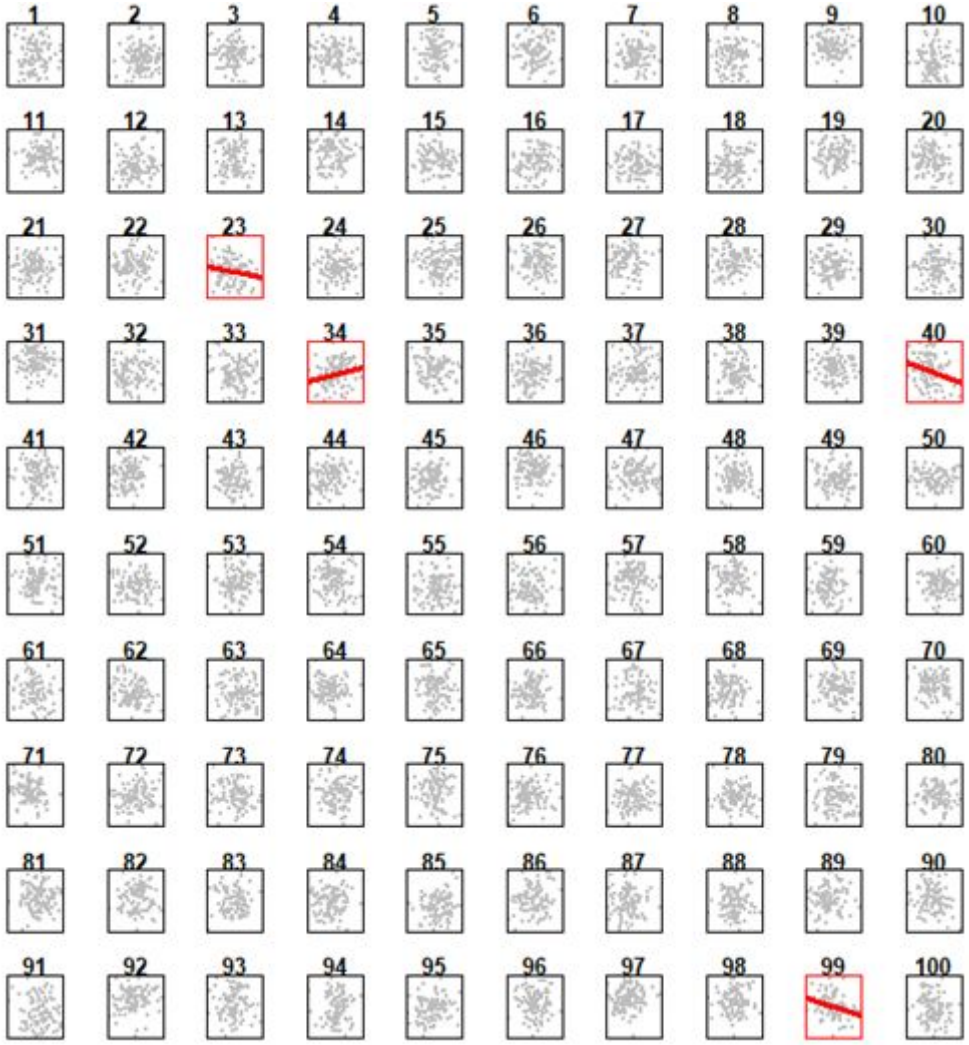
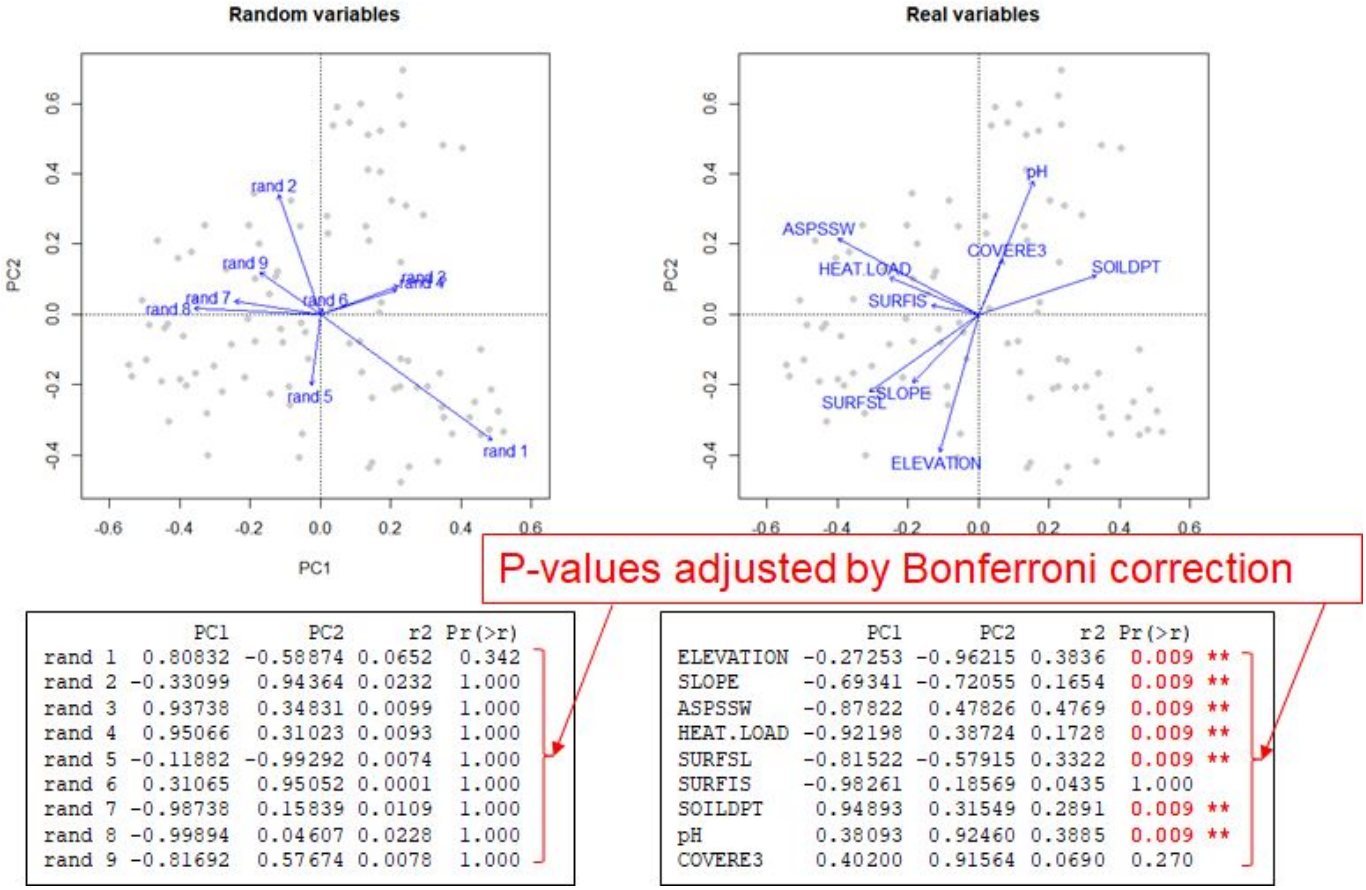


Figure 3: Multiple testing issue. I generated two random variables (normally distributed) and tested the significance of their regression with parametric F-test. I replicated this 100 times, each with newly generated random variables. Significant regressions (P < 0.05) are displayed with a red regression line. From a total of 100 analyses, four are significant at the level of 0.05 (almost 5% of all analyses).

The solution is to either avoid doing multiple tests or apply some of the corrections methods. Perhaps the best known is Bonferroni correction, which is however also very conservative (you simply multiply the resulting P-values by the overall number of tests *m* you did in the analysis, $P_{adj} = P * m$) and becomes detrimental in case that the number of tests is high, since it reduces the power of the test. Less conservative are Holm or false discovery rate (FDR) corrections. More about multiple testing issue can be found in my blog post.

In the case of example above using nine random and real supplementary variables and relating them to unconstrained ordination axes, if we apply the multiple testing correction (here Bonferroni, Fig. 4), all results in the case of random variables become insignificant (in case of the real variables, one

more result become insignificant compared to the not-corrected results). Since in this case, the test is permutational and the minimal P-value depends on the number of permutations, in case that there are many supplementary variables (and many tests), it may be necessary to increase the number of permutations to decrease the minimum P-value which can be calculated. For example, if the number of permutations is set to 199 (e.g. due to the calculation time), the minimum P-value which can be reached is P_{min} = 1/(199+1) = 0.005; if there are ten variables and the correction for multiple testing is done by Bonferroni (P-value * number of tests), the best resulting corrected P-value would be 0.005*10 = 0.05, which means that we would be unable to reject the null hypothesis on P < 0.05.



	PC1	PC2	r2	Pr(>r)
rand 1	0.80832	-0.58874	0.0652	0.342
rand 2	-0.33099	0.94364	0.0232	1.000
rand 3	0.93738	0.34831	0.0099	1.000
rand 4	0.95066	0.31023	0.0093	1.000
rand 5	-0.11882	-0.99292	0.0074	1.000
rand 6	0.31065	0.95052	0.0001	1.000
rand 7	-0.98738	0.15839	0.0109	1.000
rand 8	-0.99894	0.04607	0.0228	1.000
rand 9	-0.81692	0.57674	0.0078	1.000

	PC1	PC2	r2	Pr(>r)	
ELEVATION	-0.27253	-0.96215	0.3836	0.009	**
SLOPE	-0.69341	-0.72055	0.1654	0.009	**
ASPSSW	-0.87822	0.47826	0.4769	0.009	**
HEAT.LOAD	-0.92198	0.38724	0.1728	0.009	**
SURFSL	-0.81522	-0.57915	0.3322	0.009	**
SURFIS	-0.98261	0.18569	0.0435	1.000	
SOILDPT	0.94893	0.31549	0.2891	0.009	**
pH	0.38093	0.92460	0.3885	0.009	**
COVERE3	0.40200	0.91564	0.0690	0.270	

Figure 4: Results of multiple regression of random (left) and real (right) supplementary variables with first two axes of unconstrained ordination, with P-values adjusted by Bonferroni correction to acknowledge the multiple testing issue.